(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 February 2005 (10.02.2005)

PCT

(10) International Publication Number
WO 2005/013077 A2

(51) International Patent Classification[7]: G06F

(21) International Application Number: PCT/US2004/024351

(22) International Filing Date: 29 July 2004 (29.07.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/491,635 30 July 2003 (30.07.2003) US

(71) Applicant *(for all designated States except US)*: **UNIVERSITY OF MEDICINE AND DENTISTRY OF NEW JERSEY** [US/US]; 335 George Street, New Brunswick, NJ 08901 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **WELSH, William, J.** [US/US]; 66 Maybury Hill Road, Princeton, NJ 08540 (US). **OUYANG, Ming** [—/US]; 9 Stoecker Road, Holmdel, NJ 07733 (US). **LIOY, Paul** [US/US]; 111 Holly Street, Canford, NJ 07016 (US). **GEORGOPOULOS, Panos** [GR/US]; 70 Spring Wood Court, Princeton, NJ 08540 (US).

(74) Agents: **LICATA, Jane, Massey** et al.; Licata & Tyrell P.C., 66 E. Main Street, Marlton, NJ 08053 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

[Continued on next page]

(54) Title: SYSTEMS AND METHODS FOR MICROARRAY DATA ANALYSIS

(57) Abstract: Clustering is routinely applied in the exploratory analysis of microarray data. Missing entries arise from blemishes on the microarrays. The present invention provides a new method, and computer program and/or computer product thereof to impute missing values. The method involves the steps of clustering microarray data by partitioning the data into a select number of clusters, wherein each data point is iteratively moved from one cluster to another, until two consecutive iterations have resulted in the same partition pattern; obtaining a select number of estimates of the data in the clusters by probabilistic interference; and averaging the select number of estimates to obtain missing values in the microarray data. The method is superior to other imputation models as measured by root mean squared errors.

```
EM_ESTIMATE (μ1, ..., μK, Σ1, ..., ΣK, τ1, ..., τK, A')
{       FOR EACH ROW R OF A' WITH MISSING VALUES
{       FOR i = 1, ..., K
{       USE EM AND N(μi, Σi) TO ESTIMATE THE
        MISSING VALUES IN R.
Ri ← R WITH MISSING VALUES REPLACED BY ESTIMATES.
                }
        }
R' ← WEIGHTEDAVERAGE(R1, ..., RK).
REPLACE R IN A' BY R'.
RETURN A'.
}

K_ESTIMATE(K, A)
{       /* FIRST PART: INITIALIZATION */
B ← ROWS OF A WITHOUT MISSING VALUES.
μ1, ..., μK, Σ1, ..., ΣK, τ1, ..., τK ←
   GAUSSIAN MIXTURE CLUSTERING OF B.
A' ← EM_ESTIMATE (μ1, ..., μK, Σ1, ..., ΣK, τ1, ..., τK, A).
/* SECOND PART: ITERATION */
REPEAT
{       μ1, ..., μK, Σ1, ..., ΣK, τ1, ..., τK ←
   GAUSSIAN MIXTURE CLUSTERING OF A'.
A' ← EM_ESTIMATE (μ1, ..., μK, Σ1, ..., ΣK τ1, ..., τK, A').
} UNTIL CONVERGENCE
}

GMCimpute(S, A)
{       FOR K = 1, 2, ..., S
{       AK ← K_ESTIMATE(K, A).
}
RETURN (A1 + A2 + ... + AS) / S.
}
```

WO 2005/013077 A2

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.